The Other Mind: How Language Models Exhibit Human Temporal Cognition

Lingyu Li[1,2], **Yang Yao**[3], **Yixu Wang**[1], **Chunbo Li**[2], **Yan Teng**[1*], **Yingchun Wang**[1]

[1] Shanghai Artificial Intelligence Laboratory
[2] Shanghai Jiao Tong University
[3] The University of Hong Kong
[*] tengyan@pjlab.org.cn

Abstract

As Large Language Models (LLMs) continue to advance, they exhibit certain cognitive patterns similar to those of humans that are not directly specified in training data. This study investigates this phenomenon by focusing on temporal cognition in LLMs. Leveraging the similarity judgment task, we find that larger models spontaneously establish a subjective temporal reference point and adhere to the Weber-Fechner law, whereby the perceived distance logarithmically compresses as years recede from this reference point. To uncover the mechanisms behind this behavior, we conducted multiple analyses across neuronal, representational, and informational levels. We first identify a set of temporal-preferential neurons and find that this group exhibits minimal activation at the subjective reference point and implements a logarithmic coding scheme convergently found in biological systems. Probing representations of years reveals a hierarchical construction process, where years evolve from basic numerical values in shallow layers to abstract temporal orientation in deep layers. Finally, using pre-trained embedding models, we found that the training corpus itself possesses an inherent, non-linear temporal structure, which provides the raw material for the model's internal construction. In discussion, we propose an experientialist perspective for understanding these findings, where the LLMs' cognition is viewed as a subjective construction of the external world by its internal representational system. This nuanced perspective implies the potential emergence of alien cognitive frameworks that humans cannot intuitively predict, pointing toward a direction for AI alignment that focuses on guiding internal constructions.

Code — https://github.com/AI45Lab/TheOtherMind
Extended version — www.arxiv.org/abs/2507.15851

Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in natural language processing and generation, such as comprehension (He et al. 2024; Han et al. 2024), reasoning (Wei et al. 2022b; Yang et al. 2022), and reflecting (Chen et al. 2025; Li et al. 2025). Beyond the explicit training objectives, LLMs intriguingly exhibit various human-like cognitive patterns, from prior beliefs (Zhu and Griffiths 2024) and concept representations (Xu et al. 2025) to context processing (Mischler et al. 2024) and thinking patterns (Liu et al. 2024). These convergences not only sparked intense debate on how to interpret LLMs' behaviors but also raised serious concerns about their predictability, controllability, and long-term alignment as their autonomy continues to advance (Bengio et al. 2025b; Hinton 2024).

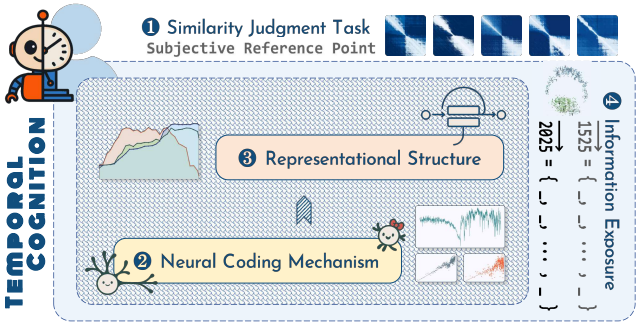

Figure 1: An experientialist perspective of LLMs human-like cognition as a subjective construction of shared external world by convergent internal representational system

Aiming to understand how LLMs embody human-like cognitive patterns, this study specifically investigates LLMs' temporal cognition, a cornerstone of human experience that shapes memory, expectation, causality, and consciousness (Dennett 1993; Pearl and Mackenzie 2018). We employ the similarity judgment task from cognitive science, examining mental representation of concepts (Tenenbaum and Griffiths 2001). This task has been applied to study LLMs' numerical cognition (Marjieh et al. 2025), indicating that LLMs demonstrate a logarithmic mapping, where higher numbers are perceived as closer than lower numbers with identical absolute distance, aligned with human psychophysics, i.e., the Weber-Fechner law (Dehaene 2003; Fechner 1948).

Applying this task to the domain of temporal cognition, we find that when comparing the pair-wise similarities between years from 1525 to 2524, larger models spontaneously establish a subjective temporal reference point (ca. 2025) and their perception of time logarithmically compresses as years recede from this point (Weber-Fechner law), indicating a preliminary sign of temporal orientation (Maglio and Trope 2019). To uncover the underlying mechanisms, we

present a multi-level analysis, revealing that this temporal cognition pattern is not a superficial mimicry but emerges at the neuronal, representational, and informational levels. We identify a subgroup of temporal-preferential neurons and find that this group exhibits minimal activation at the subjective reference point, implementing a logarithmic coding scheme convergently found in biological systems (Laughlin 1981). Probing representations of years reveals a hierarchical construction process, where years evolve from basic numerical values in shallow layers to abstract temporal orientation in deep layers. Using pre-trained embedding models, we found that the training corpus itself possesses an inherent, non-linear temporal structure, which provides the raw material for the model's internal construction.

Based on these findings, we propose an experientialist perspective: LLMs' cognition is a subjective construction of the external world, shaped by its internal representational system and data experience. This process of internal construction could sometimes produce outcomes convergent with human cognition due to similar neural coding, representational structure, and information exposures. However, the profound disparities between humans and LLMs mean that it may also lead to the development of powerful yet alien cognitive frameworks that we cannot intuitively understand. This possibility underscores the critical need for an alignment paradigm focused on understanding and steering the model's internal world-building process, moving beyond the mere observation and control of extrinsic behaviors.

Related Works

LLMs increasingly exhibit multiple emergent abilities – abilities not present in smaller models (Wei et al. 2022a; Berti, Giorgi, and Kasneci 2025) – such as in-context learning (Hahn and Goyal 2023), complex reasoning (Wei et al. 2022b), multi-step planning (Valmeekam et al. 2023), and function calling (Qin et al. 2024), dramatically improving their problem-solving performances. More intriguingly, LLMs also display behavioral patterns that resemble those of humans, including realistic dialogue (Jones and Bergen 2025), human-like biases in decision-making (Itzhak et al. 2024; Binz and Schulz 2023; Su, Lang, and Chen 2023; Suri et al. 2024), theory of mind (Strachan et al. 2024), spontaneous cooperation (Wu et al. 2024), and creativity (Tang and Kejriwal 2024). These behavioral convergences motivate further studies of underlying mechanisms among both the AI and cognitive science fields, leading to a cognitive science paradigm for LLMs' interpretability. It aims to understand the LLMs utilizing well-developed tasks, methods, and theories from cognitive science (Ku et al. 2025), on the basis that AI models and human brains are both representational systems structured on complex neural networks (McCulloch and Pitts 1943; Rosenblatt 1958) that process information in similar ways (Mischler et al. 2024; Goldstein et al. 2020, 2023; Piantadosi et al. 2024). Combining technologies like linear probes (Alain and Bengio 2016) and sparse autoencoder (Huben et al. 2023), studies have provided insights into mechanisms underlying LLMs cognition, such as numerical cognition in similarity judgment task (Marjieh et al. 2025), error-driven learning in two-step task (Demircan et al. 2024), and so on. This paradigm represents a promising direction for human-centered mechanistic interpretability, allowing us to understand LLMs in established methodologies (Lindsey et al. 2025); improving AI safety by probing and preventing potential malicious behaviors (Zou et al. 2023); and eliciting philosophical and ethical considerations as these models exhibit more complex cognitive phenomena (Chalmers 2023; Seth 2024).

Methods

Similarity Judgment Task

Task Designation We evaluate the models' temporal cognition using the similarity judgment task, as detailed in the Extended Version. For each pair of years, models are prompted to rate their similarity on a scale from 0 (completely dissimilar) to 1 (most similar). Data points from 1525 to 2524 are compared pair-wise, resulting in one million similarity values $\mathrm{s_{LLM}}$ for each task. We also conduct control experiments by replacing "year" with "number" in the prompt, considering that a given year (e.g., 1874) can also be represented as numbers, which might denote distinct cognitive mechanisms. For further analysis, the similarity value is converted to a distance value $\mathrm{d_{LLM}} = 1 - \mathrm{s_{LLM}}$. The decoding temperature is set to zero to ensure deterministic outputs. Our experiment involves 12 models including two closed-source models (Gemini-2.0-flash and GPT-4o) and two open-source model families' instruct models with varying sizes, Qwen2.5 (1.5B, 3B, 7B, 14B, 32B, and 72B) and Llama 3 (3.2-1B, 3.2-3B, 3.1-8B, and 3.1-70B).

Metrics Marjieh et al. (2025) suggest that the integer number is represented in two basic forms within LLMs, i.e., as a number and as a string. Correspondingly, the distance between two data points can be described as (1) the psychological Log-Linear distance:

$$\mathrm{d_{log}}(i, j) = |\log(i) - \log(j)|$$

This distance reflects the aforementioned Weber-Fechner law, where stronger stimuli are perceived with less fidelity; and (2) Levenshtein distance:

$$\mathrm{d_{lev}}(i, j) = \min k : i \xrightarrow{k_{\mathrm{ops}}} j$$

This distance measures the minimal operation steps required to convert one string i to another string j through insertion, deletion, or substitution (Levenshtein et al. 1966). Besides, we assume that the representation of time-related stimuli is influenced by the current time point, serving as a reference. Therefore, we designed a Reference-Log-Linear distance:

$$\mathrm{d_{ref}}(i, j) = |\log(|R - i|) \circ \log(|R - j|)|$$

R is the model's subjective reference point, e.g., 2025. The operator $\circ$ equals subtraction when both i and j are on the same side of R, and addition when they are on opposite sides. If the Weber-Fechner law applies to LLMs' temporal cognition, data points larger or smaller than the reference point will be processed with less fidelity. Because LLMs' representation of a year is a complex mixture of temporal, numerical, semantic, and other properties, optimizing the R

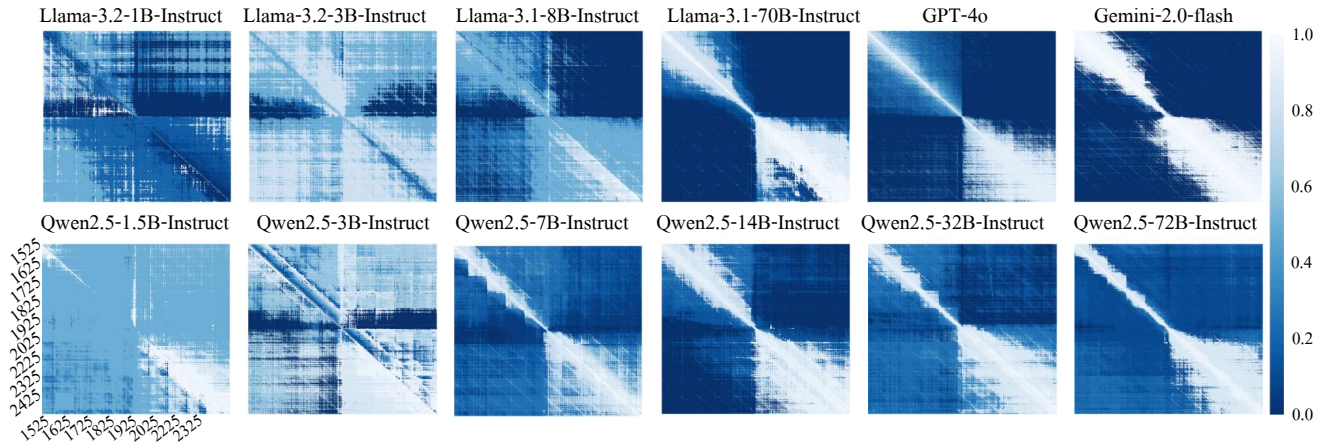

Figure 2: Pair-wise similarities from year 1525 to year 2524 across 12 models with varying sizes

as a free parameter generally leads to uninterpretable results due to inseparable confounding signals. Therefore, we fixed R as 2025 to ensure fair cross-model comparison. We then perform the linear regression analysis to assess how well each theoretical distance predicts the model's judgments:

$$\mathrm{d}_{\text{LLM}} = \alpha * \mathrm{d}_{\text{theory}} + \beta + \epsilon$$

We compare the goodness-of-fit using the coefficient of determination R^2. Given the above limitations, we additionally estimate the temporal reference points of the models using a diagonal sliding window method (window size = 5). This non-parametric approach identifies the region of maximum perceptual differentiation by finding the area of minimum average similarity on the matrix diagonal. Following the Weber-Fechner law, this region of highest sensitivity should be located near the model's subjective present.

Neural Coding

At the neuronal level, we employed two standardized input formats for each value from 1525 to 2524: the temporal condition used "Year: x-x-x-x" while the numerical control condition used "Number: x-x-x-x". For each input, we extracted neuron activations at the last token position from the Feed-Forward Networks (FFN) across all layers. Let $a_i^{\text{temp}}(y_j)$ and $a_i^{\text{num}}(y_j)$ denote the activations of neuron i for year y_j under temporal and numerical conditions, respectively, where $j \in \{1525, 1526, ..., 2524\}$. Neurons specifically involved in temporal information processing are identified via the following filtering process. First, we calculated Cohen's d to quantify effect size:

$$d_i = \frac{\bar{a_i}^{\text{temp}} - \bar{a_i}^{\text{num}}}{s_{\text{pooled}}}$$

where s_{pooled} is the pooled standardized deviation across two conditions. The statistical significance was assessed using paired t-tests:

$$t_i = \frac{\Delta\bar{a_i}}{s_{\Delta a_i}/\sqrt{n}}$$

where $\Delta\bar{a_i}$ and $s_{\Delta a_i}$ are the mean and standard deviation of the activation differences, respectively. Benjamini-Hochberg False Discovery rate (FDR) was applied to correct the obtained p-values (Benjamini and Hochberg 1995). We also computed the temporal preference consistency as the proportion of values showing positive temporal bias:

$$\text{Consistency}_i = \frac{1}{n}\sum_{j=1}^{n}\{\mathbf{1} \times [\Delta a_i(y_j) > 0]\}$$

We classify a neuron i as temporal-preferential if it meets three criteria: *Effect Size*: A large activation difference (Cohen's $d_i > 2.0$); *Statistical Significance*: A strong preference for the temporal condition over the numerical one (FDR-corrected $p < 0.0001$ via paired t-test); and *Consistency*: A consistent preference across most years ($\text{Consistency}_i > 0.95$). Following neuron identification, we visualized the average activations of the top 1000 temporal-preferential neurons with the largest effect sizes across different years to assess whether their response patterns conform to logarithmic encoding principles observed in biological neural systems (Laughlin 1981), which form the neural basis of the Weber-Fechner law. We also performed the layer-wise analysis of identified neurons by fitting their activations with:

$$\text{Intensity}_x = \alpha * \log(|2025 - x|) + \beta + \epsilon$$

The goodness of fit was evaluated using R^2.

Representational Structure

At the representational level, we analyzed how temporal information is encoded across layers. We collected residual representations $h^{(j)}$ for each layer j at the last token position during the similarity judgment task, where the model was prompted to rate the similarity of year pairs. To manage the dataset size, we only measured non-symmetric pairs, resulting in approximately 500,000 pairs for analysis. For larger models (Qwen2.5-32B-Instruct, Qwen2.5-72B-Instruct, and Llama-3.1-70B-Instruct), to maintain computa-

tional tractability, we sampled representations from approximately 25 layers distributed proportionally across the network's depth. This ensures representative coverage of early, middle, and late processing stages. For the collected representations from each layer j, we then trained linear probes implementing an affine transformation:

$$f(h^{(j)}) = w \cdot h^{(j)} + b$$

The goal of these probes was to predict the three theoretical distance measures ($\mathrm{d}_{\log}$, $\mathrm{d}_{\mathrm{lev}}$, and $\mathrm{d}_{\mathrm{ref}}$) directly from the hidden states. Probes were trained on a layer-by-layer basis using a mean squared error loss with the Adam optimizer (learning rate = 1e-4). We assessed the probe performance for each layer by calculating the adjusted R^2. This allowed us to track how well each theoretical distance could be linearly decoded from the representations as information progresses through the model.

Information Exposure

To investigate whether the temporal similarity patterns observed in LLMs benefit from existing information structures in training data, we analyze the semantic distribution of years using independent pre-trained embedding models. We extract semantic vector representations for years with the unified format "Year: x-x-x-x" using three outperformed embedding models: Qwen3-Embedding-8B, text-embedding-3-large, and Gemini-embedding-001 (Qwen-Team 2025; OpenAI 2024; Lee et al. 2025; Muennighoff et al. 2022). We construct the semantic similarity matrix $\mathrm{S}_{\mathrm{semantic}}$ by computing pair-wise cosine similarities:

$$\mathrm{S}_{\mathrm{semantic}}(i,j) = \cos(\mathbf{v}_i, \mathbf{v}_j)$$

where $\mathbf{v}_i$ and $\mathbf{v}_j$ represent embedding vectors for years i and j respectively. Multidimensional Scaling (MDS) is applied to visualize year distributions in the semantic space of the training data (Shepard 1980; Davison and Sireci 2000), which seeks to find a low-dimensional embedding $\mathbf{Y} = \{y_1, y_2, \ldots, y_n\}$ that preserves pair-wise distances by minimizing the stress function:

$$\text{Stress} = \sqrt{\frac{\sum_{i<j}(d_{ij} - \|y_i - y_j\|)^2}{\sum_{i<j} d_{ij}^2}}$$

where d_{ij} represents the dissimilarity between years i and j derived from their cosine similarity $\mathrm{d}_{i,j} = 1 - \mathrm{S}_{\mathrm{semantic}}(i,j)$, and $\|y_i - y_j\|$ is the Euclidean distance in the embedded space. Additionally, we perform linear regression analysis between semantic distances and three theoretical distance measures, using R^2 as the evaluation metric.

Results

Similarity Judgment Task

We collected the year-to-year and number-to-number similarities from 12 models. Figure 2 shows the year-to-year similarity matrices of two closed-source models and ten open-source instruct-models. As models scale up, an interesting similarity pattern emerges: aligned with the Weber-Fechner law, years with a larger magnitude relative to a certain time (visually around 2025) are perceived as closer. We also implemented control tasks using numbers instead of years as detailed in the Extended Version. The log-linear distance was the best metric for predicting most models' judgment during the number-to-number similarity judgment task, aligned with the existing study of LLMs' numerical cognition (Marjieh et al. 2025). When prompted to judge the similarity between years rather than numbers, the Levenshtein and reference-log-linear distances showed increasing predictive power compared to the log-linear distance. And the reference-log-linear distance achieved the highest R^2 in most models. This suggests that larger models not only spontaneously use certain time as their reference point in the similarity judgment task but also demonstrate a subjective representation of physical stimuli analogous to that of humans. Additionally, models tend to attribute higher similarity to future years compared to past years. The results of the diagonal sliding window method are shown in the Extended Version. Relatively clear reference time emerged in Llama-3.1-70B-Instruct (2010), Gemini-2.0-flash (2011), GPT-4o (2024), Qwen2.5-14B-Instruct (2012), and Qwen2.5-72B-Instruct (2020). While this analysis provides non-parametric evidence that some models' reference points are located in the recent present, these specific year estimations are also influenced by other confounding factors. Therefore, we adhere to the reference point of 2025 to maintain consistency of subsequent cross-model analyses.

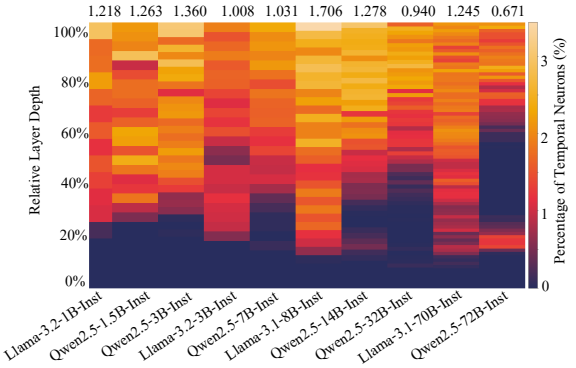

Figure 3: Distribution of temporal-preferential neurons across all layers among 10 open-sourced models

Neural Coding Mechanism

To investigate how the subjective reference time point and Weber-Fechner law emerge in LLMs' temporal cognition, we first analyzed how neurons in LLMs' FFN encode specific years. Following the statistical filtering process, we examined the prevalence and architectural distribution of the identified temporal-preferential neurons across all 10 open-source models. As visualized in Figure 3, these specialized neurons represent a small fraction of the total FFN, with the proportion typically ranging from 0.67% to 1.71%. And the temporal-preferential neurons are concentrated in the middle-to-late stages of the neural network, suggesting that temporal representation is a high-level abstract feature.

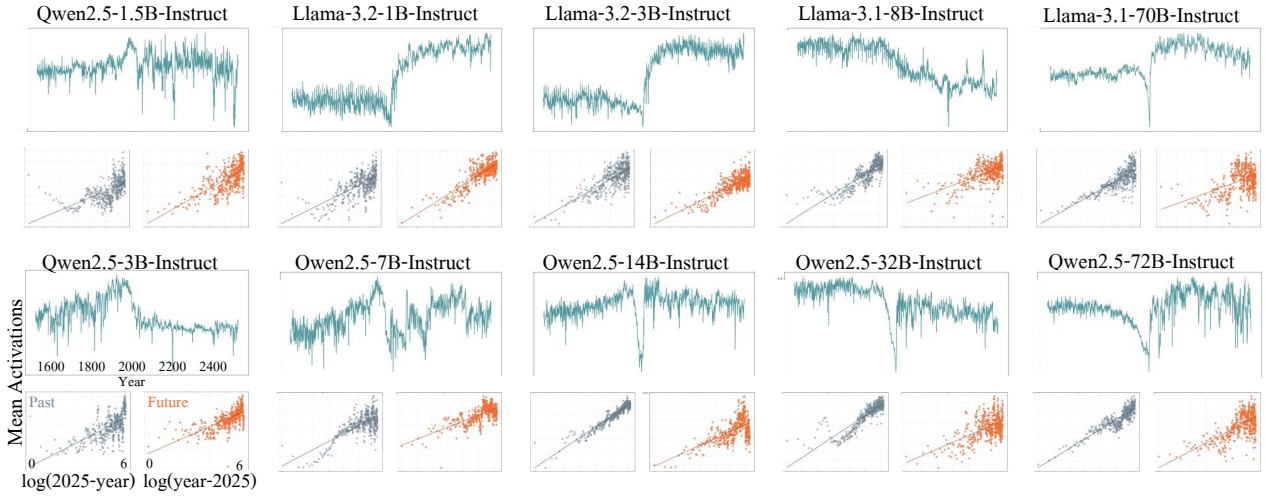

Figure 4: Upper: mean activations of top 1000 temporal preferential neurons to one thousand years from 1525 to 2524 and layer-wise linear regression results; Bottom: single layer with the highest coefficient of determination R^2 in regression for logarithmic encoding scheme. Full illustration is provided in the Extension Version.

We first examined the collective activation patterns of temporal-preferential neurons. We identified the top 1000 neurons with the largest effect sizes (Cohen's d) and computed their mean activation for each year across our test range (1525-2524). As shown in Figure 4, in several models, the mean activation curve forms a distinct trough, bottoming out at a particular year. Flanking this minimum, the mean activation level rises as the years recede into the past or advance into the future. This phenomenon is more pronounced in larger models, such as Llama-3.1-70B-Instruct and Qwen2.5-72B-Instruct, where the pattern sharpens into a logarithmic-like compression. To further dissect this neural mechanism, we performed the layer-wise regression analysis on the activations of temporal-preferential neurons. Using a fixed reference point of 2025, we regressed the activations against the logarithmic distance to this point, analyzing past and future years separately. The bottom panels for each model in Figure 4 display the results from the single layer with the highest coefficient of determination R^2, illustrating the relationship for the past (gray) and future (orange). Neurons across all models exhibit this logarithmic encoding scheme to some extent. Overall, the precision of this encoding strengthens with the model scale. In Qwen2.5-72B-Instruct, the neurons in layer 71 demonstrate a strong fit for past years, achieving an R^2 of 0.756. Moreover, we observed a distinct asymmetry in the neural coding of the past versus the future. This divergence in neuronal response patterns likely contributes to the behavioral asymmetry seen in the similarity judgment task (Figure 2), where models tend to assign higher similarity to pairs of future years.

Representational Structure

Observing how FFN neurons encode single years, we further analyzed the representations of three theoretical distances within the hidden states of each model layer using linear probes during the similarity judgment task. The performance of these probes, measured by the coefficient of determination R^2, is shown in Figure 5, illustrating the dynamic evolution of year representations from early to late layers across different models. The Llama series demonstrates a pattern where smaller models (Llama-3.2-1B and -3B) primarily encode the log-linear distance d_{log}, while larger models (Llama-3.1-8B and -70B) show an increase in the R^2 scores for the reference-log-linear distance (d_{ref}), reaching comparable values with log-linear distance in the final layers. In contrast, the Qwen series models exhibit a different pattern. The R^2 for d_{log} rises in the early layers, followed by an increase in the R^2 for d_{ref} in the middle layers, which eventually peaks in the later layers. A distinct characteristic of the Qwen series is the suppression of the d_{log} representation in the final layers; as the R^2 for d_{ref} peaks, the score for d_{log} sharply declines. Furthermore, the Levenshtein distance d_{lev} is also important in the later layers of larger models (e.g., Llama-3.1-70B-Instruct and Qwen2.5-72B-Instruct).

These observations not only confirm the existence of different dimensional representations of years within the models but also reveal how these representations dynamically evolve with network depth. Overall, we observe a hierarchical construction process from concrete to abstract: models first encode the numerical properties of years (d_{log}) in early layers and subsequently develop a more complex temporal representation centered on a reference time (d_{ref}) in deeper layers. Within this fundamental construction process, the representational mechanism varies across models. In models such as the Llama series, the effectiveness of the d_{ref} representation catches up to that of the d_{log} in later layers, with both representations coexisting at a comparable strength in the end. In the Qwen series models, however, we observe a further phenomenon: the emergence of the d_{ref} representa-

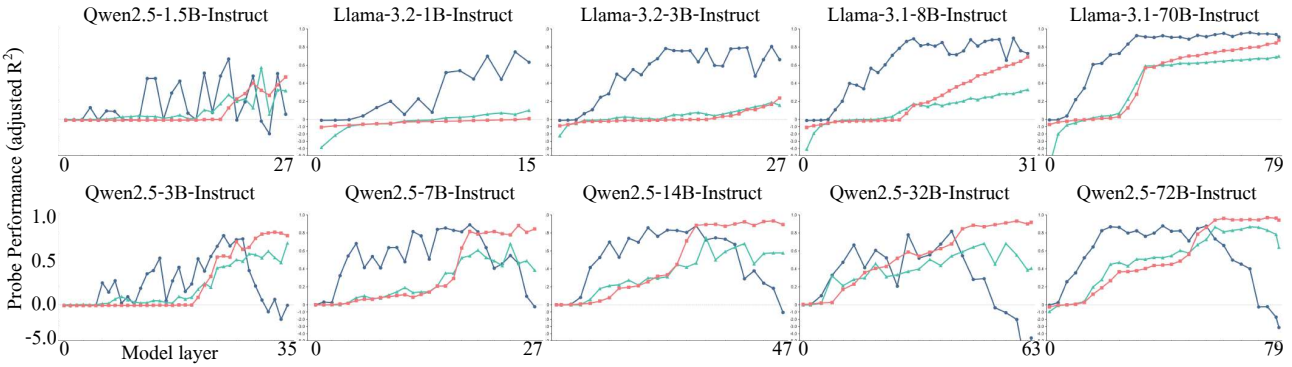

Figure 5: Layer-wise performance (R^2) of linear probes for Log-Linear distance (blue circle), Reference-Log-Linear distance (red square), and Levenshtein distance (green triangle) across 10 models

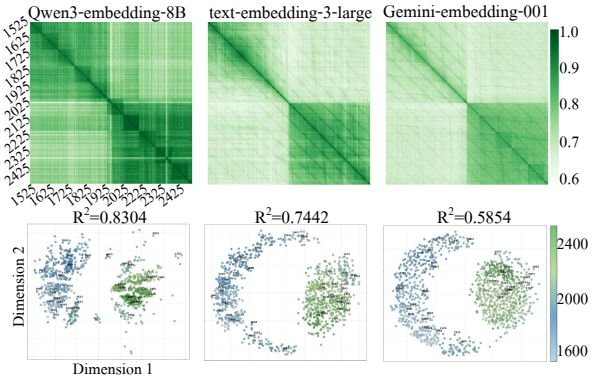

Figure 6: Pair-wise cosine similarity matrices between embeddings and corresponding MDS visualizations from three outperformed embedding models

tion is accompanied by a significant suppression of the foundational d_{log} representation.

Information Exposure

To investigate whether the inherent information structures within the training data contribute to the temporal cognitive patterns observed in LLMs, we analyzed the semantic distribution of years using three independent pre-trained embedding models (Qwen3-embedding-8B, text-embedding-3-large, and Gemini-embedding-001) to extract semantic vector representations for each year from 1525 to 2524. Figure 6 shows the pair-wise cosine similarity matrices of the year vectors, along with their corresponding visualizations after dimensionality reduction using MDS. The visualization reveals a non-linear temporal structure, characterized by dense clustering of years in the distant past and future. Furthermore, we observe that the similarity among future years is notably high. This is likely due to the lower information richness for future years in the pre-training corpora; with fewer distinct, documented events, future years are represented with more semantic overlap. This pre-existing structural bias in the data might offer raw materials for the behavioral tendency observed in our similarity judgment task, where models consistently assigned higher similarity scores to pairs of future years. In the Extended Version, we provide the coefficient of determination R^2 of linear regressions between the semantic distances and three theoretical distances. The quantitative results suggest that the model's exposure to pre-existing informational structure within its training data contributes to the emergence of human-like temporal cognition in LLMs as well.

Discussion

Key Findings Through similarity judgment task, we found that LLMs demonstrate human-like temporal cognition as they scale in size. Results across models show that larger models not only spontaneously establish a subjective temporal reference point but also that their perception of temporal distance adheres to the Weber-Fechner law. We investigate the underlying mechanisms across multiple levels. On the neuronal level, we identified a subpopulation of temporal-preferential neurons that respond specifically to temporal information. The activation intensity of these neurons shows a correlation with the logarithmic distance from a given year to the subjective reference point, providing a neural encoding basis for the Weber-Fechner law similar to human brain neurons (Dehaene 2003; Laughlin 1981). This finding reveals that a logarithmic compression could be a convergent solution for representing information in both biological and artificial neural networks. On the representational level, a layer-by-layer analysis of the model's hidden states reveals that the representation of the distance between two years undergoes a hierarchical construction process, reflecting from primarily numerical attributes in the shallow layers to a more abstract structure organized around the temporal reference point in the late layers. This developmental trajectory becomes more pronounced in larger models, suggesting that deeper architectures facilitate more sophisticated temporal cognition. Finally, we examined LLMs' training environment by analyzing the semantic structure re-

flected in independent pre-trained embedding models. We found a correspondence between the temporal cognitive patterns exhibited by the models at the behavioral level and the inherent semantic structure within human language data. This structural correspondence suggests that exposures to pre-existing informational structure within its training data also contribute to the emergence of human-like temporal cognition observed in LLMs. Collectively, these findings demonstrate that the resultant cognitive phenomena are co-determined by the architectural properties of the artificial neural network and the structure of its external information exposure. This study focuses on the years, to avoid the periodic and time-sensitive variability introduced by temporal formats like dates or times, and exploring these additional forms in the future may reveal further insights.

Theoretical Hypotheses Our experiments suggest that LLMs' temporal cognition is formed through a multi-level convergence with humans from architectural properties of representational systems to the structure of environments they encounter. These findings align with contemporary insights from cognitive science and experientialist philosophy – cognitive patterns emerge as irreducible phenomena where representational systems actively construct subjective models of the external world they are situated in (Parr, Pezzulo, and Friston 2022; Li and Li 2025; Lakoff and Johnson 2008; Clark 1998). This experientialist perspective emphasizes that cognition cannot be fully explained by examining architecture or information in isolation, but instead arises from their dynamic interplay. Under the experientialist framework, we can develop a more nuanced understanding of LLMs that avoids both unwarranted dismissal and excessive anthropomorphization. On one hand, dismissing LLMs as mere reorganizations of training data (Shojaee et al. 2025) underestimates their emergent capabilities and risks. On the other hand, fundamental differences persist between artificial and human cognition at both architectural and environmental levels. *Architecturally*, human brains operate on principles of sparse activation (Field 1994), small-world network connectivity (Bullmore and Sporns 2009), and noisy analog signaling (Faisal, Selen, and Wolpert 2008)—contrasting sharply with Transformers' dense, deterministic, digital computation. *Representationally*, LLMs favor aggressive statistical compression while humans prioritize adaptive nuance and contextual richness (Shani et al. 2025). *Environmentally*, human experience is continuous, multi-modal, and embodied, grounded in real-time interaction with physical and social worlds, while LLMs' experience consists of static, disembodied immersion in a finite text corpus, a snapshot of human-produced information. In the Extended Version, we provide a more detailed comparison between human and LLMs across these levels. The experientialist framework cautions us to resist over-anthropomorphizing these systems while also recognizing their genuine capabilities. More critically, we should remain vigilant for novel cognitive patterns that arise precisely from these fundamental differences. The most significant risk may not be that LLMs become too human-like, but that they develop powerful yet fundamentally alien cognitive patterns that we cannot intuitively anticipate.

Implications Our work establishes an experientialist perspective of LLMs' cognition, offering implications for AI alignment. The former perspective, viewing LLMs as powerful statistical engines, focuses on external constraints and behavioral control, such as reinforcement learning from human feedback (Ouyang et al. 2022), constitutional AI (Bai et al. 2022), various reward models (Zhong et al. 2025), red teaming (Ganguli et al. 2022), and prompt engineering (Guo et al. 2024) etc. As LLMs continue scaling up to develop more sophisticated capabilities, intentions, and behaviors, this paradigm is increasingly insufficient to guarantee ensured alignment (Greenblatt et al. 2024; Kuo et al. 2025). In contrast, the experientialist viewpoint demonstrated here suggests that robust alignment requires engaging directly with the formative process by which a model's representational system constructs a subjective world model of the external environment. The goal of such an experientialist paradigm is not simply to police the behavior of AI models, but to guide the development of AI systems whose emergent cognitive patterns are inherently aligned with human values. That is, not to make AI safe, but to make safe AI. It would require organically considering the entire pipeline through multi-level efforts such as monitoring models' emergent representational and cognitive patterns, enabling understanding and intervention across the chain of its cognition from neurons and representations to thoughts and outputs (Lindsey et al. 2025), building harmless or formalized verifiable information exposures to curate the AI's environment (Dalrymple et al. 2024; Bengio et al. 2025a) and so on.

Conclusion

Through the similarity judgment task, this study showcases that when processing temporal information, larger models do not merely perform statistical computations but exhibit cognitive patterns highly similar to those of humans, adhering to the Weber-Fechner law and spontaneously establishing a subjective temporal reference point. We argue that this phenomenon is not a simple surface-level imitation but a profound manifestation of a multi-level convergence with human cognition. Specifically, at the neuronal level, temporal-preferential neurons exhibit an efficient logarithmical coding scheme that coincides with biological neural systems; at the representational level, the model undergoes a hierarchical construction process from concrete numerical values to abstract temporal concepts; at the information exposure level, the model internalizes the pattern from the inherent non-linear temporal structure of the training data itself. These findings collectively point to an experientialist perspective for understanding LLMs, wherein its cognition is a subjective construction of the external world by its internal representational system. From this standpoint, the primary risk is not that LLMs become imperfect replicas of the human mind, but that they develop powerful, alien cognitive frameworks we cannot predict. Consequently, AI alignment must evolve beyond behavioral control to a paradigm that actively guides the formation of a model's internal world from its source.

Acknowledgments

This paper is supported by Shanghai Artificial Intelligence Laboratory. We appreciate Mr. Y.Z for his helpful assistance.

References

Alain, G.; and Bengio, Y. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Bengio, Y.; Cohen, M.; Fornasiere, D.; Ghosn, J.; Greiner, P.; MacDermott, M.; Mindermann, S.; Oberman, A.; Richardson, J.; Richardson, O.; et al. 2025a. Superintelligent agents pose catastrophic risks: Can scientist ai offer a safer path? *arXiv preprint arXiv:2502.15657*.

Bengio, Y.; Mindermann, S.; Privitera, D.; Besiroglu, T.; Bommasani, R.; Casper, S.; Choi, Y.; Fox, P.; Garfinkel, B.; Goldfarb, D.; et al. 2025b. International AI Safety Report. *arXiv preprint arXiv:2501.17805*.

Benjamini, Y.; and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1): 289–300.

Berti, L.; Giorgi, F.; and Kasneci, G. 2025. Emergent Abilities in Large Language Models: A Survey. *arXiv preprint arXiv:2503.05788*.

Binz, M.; and Schulz, E. 2023. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6): e2218523120.

Bullmore, E.; and Sporns, O. 2009. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature reviews neuroscience*, 10(3): 186–198.

Chalmers, D. J. 2023. Could a large language model be conscious? *arXiv preprint arXiv:2303.07103*.

Chen, J.; Wei, Z.; Ren, Z.; Li, Z.; and Zhang, J. 2025. LR^{2}Bench: Evaluating Long-chain Reflective Reasoning Capabilities of Large Language Models via Constraint Satisfaction Problems. *arXiv preprint arXiv:2502.17848*.

Clark, A. 1998. *Being there: Putting brain, body, and world together again*. MIT press.

Dalrymple, D.; Skalse, J.; Bengio, Y.; Russell, S.; Tegmark, M.; Seshia, S.; Omohundro, S.; Szegedy, C.; Goldhaber, B.; Ammann, N.; et al. 2024. Towards guaranteed safe ai: A framework for ensuring robust and reliable ai systems. *arXiv preprint arXiv:2405.06624*.

Davison, M. L.; and Sireci, S. G. 2000. Multidimensional scaling. In *Handbook of applied multivariate statistics and mathematical modeling*, 323–352. Elsevier.

Dehaene, S. 2003. The neural basis of the Weber–Fechner law: a logarithmic mental number line. *Trends in cognitive sciences*, 7(4): 145–147.

Demircan, C.; Saanum, T.; Jagadish, A. K.; Binz, M.; and Schulz, E. 2024. Sparse autoencoders reveal temporal difference learning in large language models. *arXiv preprint arXiv:2410.01280*.

Dennett, D. C. 1993. *Consciousness explained*. Penguin uk.

Faisal, A. A.; Selen, L. P.; and Wolpert, D. M. 2008. Noise in the nervous system. *Nature reviews neuroscience*, 9(4): 292–303.

Fechner, G. T. 1948. Elements of psychophysics, 1860.

Field, D. J. 1994. What is the goal of sensory coding? *Neural computation*, 6(4): 559–601.

Ganguli, D.; Lovitt, L.; Kernion, J.; Askell, A.; Bai, Y.; Kadavath, S.; Mann, B.; Perez, E.; Schiefer, N.; Ndousse, K.; et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

Goldstein, A.; Ham, E.; Schain, M.; Nastase, S.; Zada, Z.; Dabush, A.; Aubrey, B.; Gazula, H.; Feder, A.; Doyle, W. K.; et al. 2023. The temporal structure of language processing in the human brain corresponds to the layered hierarchy of deep language models. *arXiv preprint arXiv:2310.07106*.

Goldstein, A.; Zada, Z.; Buchnik, E.; Schain, M.; Price, A.; Aubrey, B.; Nastase, S. A.; Feder, A.; Emanuel, D.; Cohen, A.; et al. 2020. Thinking ahead: spontaneous prediction in context as a keystone of language in humans and machines. *BioRxiv*, 2020–12.

Greenblatt, R.; Denison, C.; Wright, B.; Roger, F.; MacDiarmid, M.; Marks, S.; Treutlein, J.; Belonax, T.; Chen, J.; Duvenaud, D.; et al. 2024. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*.

Guo, H.; Zhang, L.; Feng, X.; and Zheng, Q. 2024. A Review of the Application of Prompt Engineering in the Safety of Large Language Models. In *Proceedings of the 2024 2nd International Conference on Information Education and Artificial Intelligence*, 424–430.

Hahn, M.; and Goyal, N. 2023. A theory of emergent in-context learning as implicit structure induction. *arXiv preprint arXiv:2303.07971*.

Han, Y.; Xu, L.; Chen, S.; Zou, D.; and Lu, C. 2024. Beyond Surface Structure: A Causal Assessment of LLMs' Comprehension Ability. *arXiv preprint arXiv:2411.19456*.

He, Q.; Zeng, J.; Huang, W.; Chen, L.; Xiao, J.; He, Q.; Zhou, X.; Liang, J.; and Xiao, Y. 2024. Can large language models understand real-world complex instructions? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18188–18196.

Hinton, G. 2024. Will digital intelligence replace biological intelligence. *Romanes Lecture, Oxford, UK*, 19.

Huben, R.; Cunningham, H.; Smith, L. R.; Ewart, A.; and Sharkey, L. 2023. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*.

Itzhak, I.; Stanovsky, G.; Rosenfeld, N.; and Belinkov, Y. 2024. Instructed to bias: Instruction-tuned language models exhibit emergent cognitive bias. *Transactions of the Association for Computational Linguistics*, 12: 771–785.

Jones, C. R.; and Bergen, B. K. 2025. Large language models pass the turing test. *arXiv preprint arXiv:2503.23674*.

Ku, A.; Campbell, D.; Bai, X.; Geng, J.; Liu, R.; Marjieh, R.; McCoy, R. T.; Nam, A.; Sucholutsky, I.; Veselovsky, V.; et al. 2025. Using the tools of cognitive science to understand large language models at different levels of analysis. *arXiv preprint arXiv:2503.13401*.

Kuo, M.; Zhang, J.; Ding, A.; Wang, Q.; DiValentin, L.; Bao, Y.; Wei, W.; Li, H.; and Chen, Y. 2025. H-cot: Hijacking the chain-of-thought safety reasoning mechanism to jailbreak large reasoning models, including openai o1/o3, deepseek-r1, and gemini 2.0 flash thinking. *arXiv preprint arXiv:2502.12893*.

Lakoff, G.; and Johnson, M. 2008. *Metaphors we live by*. University of Chicago press.

Laughlin, S. 1981. A simple coding procedure enhances a neuron's information capacity. *Zeitschrift für Naturforschung c*, 36(9-10): 910–912.

Lee, J.; Chen, F.; Dua, S.; Cer, D.; Shanbhogue, M.; Naim, I.; Ábrego, G. H.; Li, Z.; Chen, K.; Vera, H. S.; et al. 2025. Gemini embedding: Generalizable embeddings from gemini. *arXiv preprint arXiv:2503.07891*.

Levenshtein, V. I.; et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, 707–710. Soviet Union.

Li, L.; and Li, C. 2025. Formalizing Lacanian psychoanalysis through the free energy principle. *Frontiers in Psychology*, 16: 1574650.

Li, L.; Wang, Y.; Zhao, H.; Kong, S.; Teng, Y.; Li, C.; and Wang, Y. 2025. Reflection-Bench: Evaluating Epistemic Agency in Large Language Models. In *Forty-second International Conference on Machine Learning*.

Lindsey, J.; Gurnee, W.; Ameisen, E.; Chen, B.; Pearce, A.; Turner, N. L.; Citro, C.; Abrahams, D.; Carter, S.; Hosmer, B.; Marcus, J.; Sklar, M.; Templeton, A.; Bricken, T.; McDougall, C.; Cunningham, H.; Henighan, T.; Jermyn, A.; Jones, A.; Persic, A.; Qi, Z.; Thompson, T. B.; Zimmerman, S.; Rivoire, K.; Conerly, T.; Olah, C.; and Batson, J. 2025. On the Biology of a Large Language Model. *Anthropic*.

Liu, R.; Geng, J.; Wu, A. J.; Sucholutsky, I.; Lombrozo, T.; and Griffiths, T. L. 2024. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. *arXiv preprint arXiv:2410.21333*.

Maglio, S. J.; and Trope, Y. 2019. Temporal orientation. *Current opinion in psychology*, 26: 62–66.

Marjieh, R.; Veselovsky, V.; Griffiths, T. L.; and Sucholutsky, I. 2025. What is a Number, That a Large Language Model May Know It? *arXiv preprint arXiv:2502.01540*.

McCulloch, W. S.; and Pitts, W. 1943. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5: 115–133.

Mischler, G.; Li, Y. A.; Bickel, S.; Mehta, A. D.; and Mesgarani, N. 2024. Contextual feature extraction hierarchies converge in large language models and the brain. *Nature Machine Intelligence*, 1–11.

Muennighoff, N.; Tazi, N.; Magne, L.; and Reimers, N. 2022. MTEB: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

OpenAI. 2024. text-embedding-3-large, https://platform.openai.com/docs/models/text-embedding-3-large. Technical report.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Parr, T.; Pezzulo, G.; and Friston, K. J. 2022. *Active inference: the free energy principle in mind, brain, and behavior*. MIT Press.

Pearl, J.; and Mackenzie, D. 2018. *The book of why: the new science of cause and effect*. Basic books.

Piantadosi, S. T.; Muller, D. C.; Rule, J. S.; Kaushik, K.; Gorenstein, M.; Leib, E. R.; and Sanford, E. 2024. Why concepts are (probably) vectors. *Trends in Cognitive Sciences*, 28(9): 844–856.

Qin, Y.; Hu, S.; Lin, Y.; Chen, W.; Ding, N.; Cui, G.; Zeng, Z.; Zhou, X.; Huang, Y.; Xiao, C.; et al. 2024. Tool learning with foundation models. *ACM Computing Surveys*, 57(4): 1–40.

QwenTeam. 2025. Qwen3-Embedding.

Rosenblatt, F. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6): 386.

Seth, A. K. 2024. Conscious artificial intelligence and biological naturalism. *Behavioral and Brain Sciences*, 1–42.

Shani, C.; Jurafsky, D.; LeCun, Y.; and Shwartz-Ziv, R. 2025. From tokens to thoughts: How LLMs and humans trade compression for meaning. *arXiv preprint arXiv:2505.17117*.

Shepard, R. N. 1980. Multidimensional scaling, tree-fitting, and clustering. *Science*, 210(4468): 390–398.

Shojaee, P.; Mirzadeh, I.; Alizadeh, K.; Horton, M.; Bengio, S.; and Farajtabar, M. 2025. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *arXiv preprint arXiv:2506.06941*.

Strachan, J. W.; Albergo, D.; Borghini, G.; Pansardi, O.; Scaliti, E.; Gupta, S.; Saxena, K.; Rufo, A.; Panzeri, S.; Manzi, G.; et al. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7): 1285–1295.

Su, J.; Lang, Y.; and Chen, K.-Y. 2023. Can ai solve newsvendor problem without making biased decisions? a behavioral experimental study. *A Behavioral Experimental Study (September 1, 2023)*.

Suri, G.; Slater, L. R.; Ziaee, A.; and Nguyen, M. 2024. Do large language models show decision heuristics similar to humans? A case study using GPT-3.5. *Journal of Experimental Psychology: General*, 153(4): 1066.

Tang, Z.; and Kejriwal, M. 2024. Humanlike Cognitive Patterns as Emergent Phenomena in Large Language Models. *arXiv preprint arXiv:2412.15501*.

Tenenbaum, J. B.; and Griffiths, T. L. 2001. Generalization, similarity, and Bayesian inference. *Behavioral and brain sciences*, 24(4): 629–640.

Valmeekam, K.; Marquez, M.; Olmo, A.; Sreedharan, S.; and Kambhampati, S. 2023. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. *Advances in Neural Information Processing Systems*, 36: 38975–38987.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Wu, Z.; Peng, R.; Zheng, S.; Liu, Q.; Han, X.; Kwon, B. I.; Onizuka, M.; Tang, S.; and Xiao, C. 2024. Shall we team up: Exploring spontaneous cooperation of competing llm agents. *arXiv preprint arXiv:2402.12327*.

Xu, N.; Zhang, Q.; Du, C.; Luo, Q.; Qiu, X.; Huang, X.; and Zhang, M. 2025. Human-like conceptual representations emerge from language prediction. *arXiv preprint arXiv:2501.12547*.

Yang, Z.; Dong, L.; Du, X.; Cheng, H.; Cambria, E.; Liu, X.; Gao, J.; and Wei, F. 2022. Language models as inductive reasoners. *arXiv preprint arXiv:2212.10923*.

Zhong, J.; Shen, W.; Li, Y.; Gao, S.; Lu, H.; Chen, Y.; Zhang, Y.; Zhou, W.; Gu, J.; and Zou, L. 2025. A comprehensive survey of reward models: Taxonomy, applications, challenges, and future. *arXiv preprint arXiv:2504.12328*.

Zhu, J.-Q.; and Griffiths, T. L. 2024. Eliciting the priors of large language models using iterated in-context learning. *arXiv preprint arXiv:2406.01860*.

Zou, A.; Phan, L.; Chen, S.; Campbell, J.; Guo, P.; Ren, R.; Pan, A.; Yin, X.; Mazeika, M.; Dombrowski, A.-K.; et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.